

METAFORA O VÝZKUMU
MULTIFAKTORIÁLNĚ PODMÍNĚNÝCH CHOROB

Petr Vopěnka, MFF UK, Praha

Dan Pokorný, Ulm

Jiří Šimek, UK, Praha

1. O MULTIFAKTORIÁLNĚ PODMÍNĚNÝCH CHOROBÁCH

V posledních desetiletích je věnována velká pozornost výzkumu "multifaktoriálně podmíněných chorob", se složitou etiologií, k nimž řadíme arteriosklerózu, zhoubné nádory, vředy žaludku a dvanácterníku atd. Pro tyto choroby bylo řadou výzkumů odhaleno mnoho rizikových faktorů, např. pro infarkt myokardu rizikové faktory hypertenze, stresující podmínky, kouření atd. Tyto výsledky byly získány převážně ve studiích dvou typů:

Klinická studie: Jsou srovnávány skupina nemocných a skupina kontrolních osob. Osoby mohou být vyšetřovány velmi podrobně a pečlivě; bohužel však rozsahy souborů bývají relativně malé.

Perspektivní studie: Je vyšetřován rozsáhlý soubor osob, vybraných u určité populace a tento soubor je podrobován přešetření v pravidelných intervalech v období řekněme deseti let. Tak je umožněno jemnější posuzování vztahu mezi příčinami a následky. Velký rozsah souboru /např. framinghamská studie na 5 209 lidech/ přináší na jedné straně možnost hlubších analýz, na straně druhé však vede k využívání méně přesných epidemiologických observačních technik.

Výsledky klinických i prospektivních studií /někdy souhlasné, někdy si odporující/ podstatně pomohly k lepšímu porozumění patogenetických podmínek nemocí se složitou etiologií. Umění analýzy dat hrálo podstatnou roli v těchto studiích ...

Přes všechny tyto úspěchy však postrádáme přesnější představy patogenetických mechanismů nemocí. Opětovně jsou podnikány pokusy konstruovat pro danou chorobu patogenetické řetězce, vedoucí od různých stavů či situací až k počátku choroby. Ať těchto řetězců máme navrhuto sebevíce, vždy nalézáme nové reálně existující protipříklady, vymykající se všem navrženým řetězcům. Neumíme vysvětlit, proč každý pacient má méně či více specifický obraz individuálních potíží. Přes veškerý pokrok ve výzkumu multifaktoriálních chorob, naše skutečné pochopení patogenetických mechanismů se příliš nezlepšilo.

Existují známá a na symposiích již téměř zlidovělá vysvětlení popsanych těžkostí:

/1/ Vznik nemocí je podmíněn daleko více okolnostmi než jsme schopni sledovat.

- /2/ Patogenetický mechanismus je nesmírně komplikovaný.
- /3/ Patogenetický mechanismus má stochastický charakter.
- /4/ Veličiny jsou často sledovány nepřesně nebo dokonce chybně.
- /5/ Klinický obraz se mění v místě a čase.
- /6/ Rozsahy souborů jsou nedostatečné.

Souhlasíme s tím, že všechny tyto okolnosti podstatným způsobem komplikují lékařský výzkum. Máme však podezření, že matematici i lékaři někdy tiše věří, že po překonání těchto obtíží by již nic nestálo v cestě odhalení "skutečných" patogenetických mechanismů. Pokusíme se ukázat, že hypotetický badatel, který by se se všemi obtížemi vyrovnal, by zřejmě narazil na obtíže další a patrně zásadnější. K této demonstraci popíšeme výsledky tří studií určité choroby.

2. 1. STUDIE UNIVERSITY V INGEN

Byly zkoumány dvě skupiny osob: 80 zdravých a 80 nemocných osob. U každé osoby bylo vyšetřeno 50 dichotomických příznaků /ano-ne/. Z věcného hlediska se těchto 50 veličin dělilo do 5 oblastí, odpovídajících různým rovinám klinického uvažování. Uvedeme namátkou jen některé oblasti a v nich zahrnuté příznaky:

Sociální vztahy: vážné problémy v rodině,
vážné problémy v práci,

...

Fysiologické regulace: vysoký krevní tlak,
nadváha,
hypercholesterolemie,

...

atd.

Protože nás - v rámci tohoto příspěvku - zajímají spíše obecné strukturální otázky než jednotlivé konkrétní výsledky, pojmenujeme tyto oblasti formálně A, B, C, D, E a proměnné v nich obsažené A1, A2, ..., A9, A0, B,1, ..., B0, C1, ..., C0, D1, ..., D0, E1, ..., E0.

Stejně formálně pojmenujeme výzkumné středisko, kde byla studie provedena, "Universita v Ingen".

Na universitě v Ingen tedy shromáždili 80 záznamů o nemocných a 80 záznamů o "zdravých" osobách. Každý záznam obsahoval 51 dichotomických údajů, 50 veličin A1, ..., E0 a přítomnost/nepřítomnost choroby. Uvedeme alespoň tři záznamy z každé skupiny /úplný materiál je přístupný na uvedené universitě/:

```
0100100000 1011000000 1001100101 0101010000 0000010100 1
0000000000 1110000000 0110100100 0001100010 0000010011 1
1101010010 1010100101 0010000001 0000001000 0110000000 1
...
0010010001 0000001010 0010010100 0000101001 0100011010 0
0100000010 0000000000 0000000000 0000100110 1001001100 0
1101000000 1000010000 0000110001 0010000101 0101000100 0
```

Co přinesla analýza dat v Ingen? Během obvyklých prvních roků byly získány základní jednorozměrné popisy sledovaných veličin. V dalším kroku byly studovány párové vztahy souvislosti mezi rizikovými faktory a výskytem choroby. Fisherův test souvislosti na hladině 0.05 vyšel pro 11 rizikových faktorů /např. E1/ významně, pro zbývajících 39 faktorů /např. A1/ nikoli:

	A1	ne A1		E1	ne E1	
nemocní	15	65	80	20	60	80
zdraví	14	66	80	14	66	80
	29	131	160	34	126	160
	Fisher = 0.50			Fisher = 0.0091		

Celkové schéma párových souvislostí rizikových faktorů a výskytu choroby:

	1	2	3	4	5	6	7	8	9	0
A
B	.	+	++	+	.	.	+	.	.	.
C	.	.	+	+
D	+	+	.	++
E	++	.	++

- . bez významné souvislosti kladné či záporné
- + kladná souvislost, hladina 0.05
- ++ kladná souvislost, hladina 0.01

V dalším kroku kontingenční analýzy byly prozkoumány některé třírozměrné tabulky /2 faktory + nemoc/. Nebyly nalezeny žádné interakce vyššího řádu, akceptované modely obsahovaly interakce druhého řádu.

Z mnohorozměrných technik byla aplikována především postupná diskriminační analýza. V běhu této procedury bylo z 50 veličin vybráno 11 a navrženo následující klasifikační kritérium:

$$12 A7 + 17 B1 + 11 B3 + 10 B4 + 12 B5 + 15 C3 + 11 C4 + 14 D1 + 13 D4 + 20 E1 + 16 E3 \geq 35$$

Toto kritérium bylo v obou skupinách úspěšné v 73.8 procentech, tj. u 59 osob z 80 /klasifikace typu jackknife/. Všimněme si relativně velmi dobré shody mezi veličinami nalezenými v diskriminační analýze a v předchozí analýze pomocí kontingenčních tabulek. Další použité procedury /např. shluková analýza veličin/ nepřinesly výrazné pohledy. Jiné procedury, pro dichotomická data zřejmě vhodnější než diskriminační analýza, nebyly použity pro obtížnější dostupnost programů. Je ovšem otázka, zda by aplikace těchto procedur zde přinesla podstatně hlubší výsledky.

Závěry lékařského výzkumníka v Ingen /Dr. Inga/ byly následující: U 11 veličin bylo prokázáno, že jsou rizikovými faktory studované choroby, u čtyř z těchto veličin s vysokým stupněm jistoty. Je překvapivé, že oblast A nemá k chorobě prakticky žádný vztah a oblast C jen slabý a nanejvýš nepřesvědčivý. Důležité jsou naopak oblasti B /veličiny B1 až B4/, D

Podle očekávání mnohorozměrná analýza tuto strukturu potvrdila.

Závěry prof. Kinga zněly tvrdě v Ingen i Yingen: Nemoc úzce souvisí s oblastí A /což zcela přehlédli dr. Ing/, dále s oblastmi D a E. K různým nálezům v oblastech B a C, oznamovaným v Ingen a Yingen, lze říci jen tolik, že se zřejmě jedná o artefakty, způsobené ne právě nejmodernější měřicí technikou, nereprezentativností výběrových souborů a podobnými problémy, bohužel stále ještě komplikujícími lékařským výzkum.

Studie K2 a K3. Jedině prof. King z nejbohatší kingenské university byl schopen zorganizovat větší studii K2 svým rozsahem 400 nemocných a 4 000 zdravých osob řádově odpovídající framinghamské studii. Při přípravě této nákladné studie bylo s přihlédnutím k výsledkům kingenského "předvýzkumu" rozhodnuto sledovat pouze oblasti A, D, E. Výsledky studie byly novým velkým úspěchem týmu prof. Kinga:

	1	2	3	4	5	6	7	8	9	0
A	++	++	++	++	++	++	++	.	+	.
D	++	++	++	++	++	+
E	++	++	++	++	++

Poslední studie K3 s rozsahy 4 000 nemocných a 4 000 zdravých, v lékařském výzkumu normálně spíše nedosažitelnými, potvrdila tyto nálezy. Navíc byly nalezeny některé interakce vyšších řádů; např. kombinace současně se vyskytujících faktorů A1 a A2 je nebezpečnější, než by bylo možno soudit z účinků samotného faktoru A1, resp. A2.

Závěr: Ačkoli kingenské studie /pochopitelně/ nepodávají plné vysvětlení všech možných patogenetických řetězců, jasně prokázaly a popsaly role rizikových oblastí A, D, E. Názory tzv. Kingenské školy jsou široce uznávány a položily základ k modernímu lékařskému přístupu k diskutované chorobě ...

3. CHOROBA

Čtenář patrně pochopil, že naše choroba je umělá, "fiktivní" a vymyšlená autory. Patogenetický mechanismus je relativně jednoduchý. Výskyt choroby závisí pouze na 25 proměnných, totiž X1, ..., X5 v oblastech A, ..., E.

Oblast. Předpokládejme, že prvních pět příznaků oblasti X /kde X=A, ..., E/ jsou cyklicky uspořádány, tj. veličina X1 je soused jak X2, tak X5. Bud' m maximální počet cyklicky sousedících hodnot "jedna".

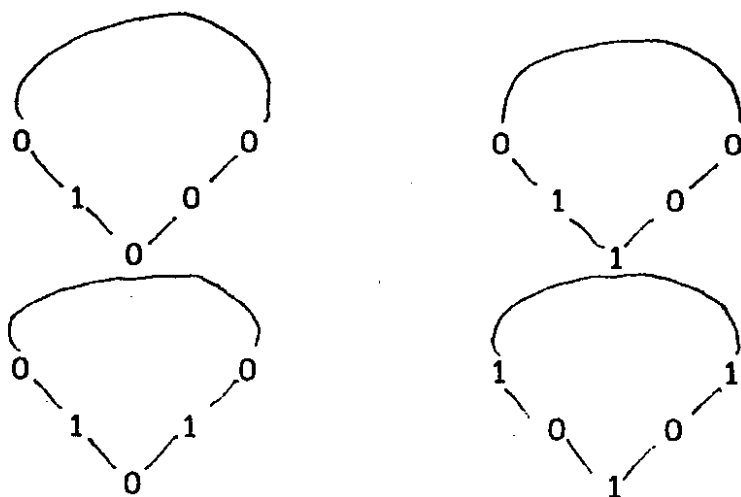
Je-li $m=0,1$ je oblast X intaktní,

je-li $m=2,3$ je oblast X patologicky poškozená,

je-li $m=4,5$ je oblast X letálně poškozená.

/Srv. obr. 1/

Obr. 1



Intaktní oblasti

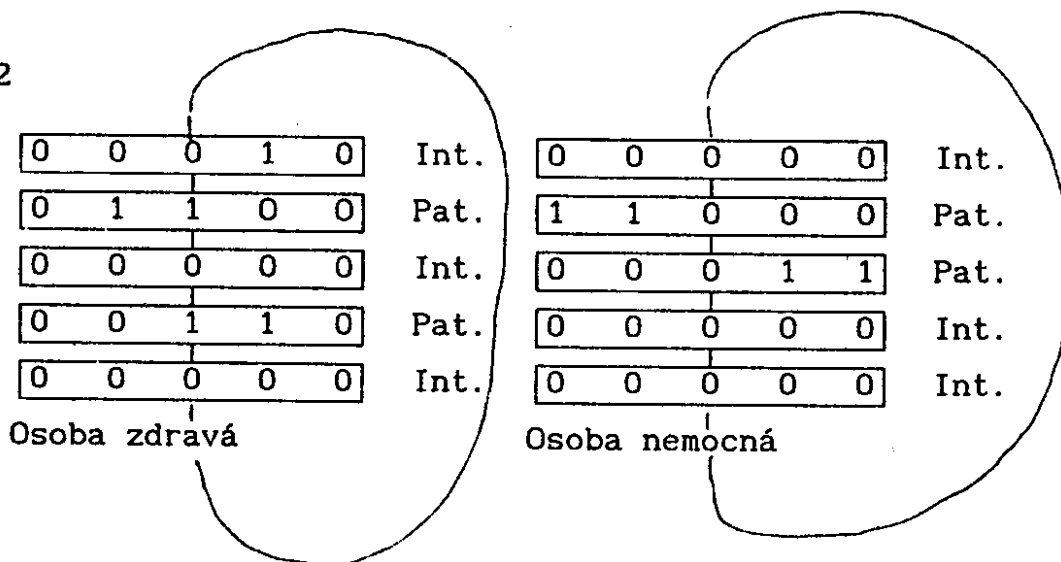
Patologicky poškozené oblasti

Osoba. Předpokládejme, že oblasti A, B, C, D, E jsou cyklicky uspořádané, tj. např. oblast E sousedí s oblastí D a s oblastí A. Pokud je kterákoli z oblastí A ... E letálně poškozená, osoba nežije. V opačném případě buď n maximální počet cyklicky sousedících patologicky poškozených oblastí.

Je-li $n=0,1$ je osoba zdravá,
 je-li $n=2,3,4,5$ je osoba nemocná,

/Srv. obr. 2/

Obr. 2



4. MODEL

Modelovali jsme situaci výzkumné studie, zahrnující: /1/ nemoc, vyskytující se v populaci a /2/ data-analytický popis jejího výskytu.

Populace. Každá veličina, A_1, \dots, E_0 má binomiální rozdělení, kde "nebezpečná" hodnota "jedna" má pravděpodobnost $p=0.2$. Veličiny z různých oblastí /např. A_2 a C_7 / jsou nezávislé. Dvě veličiny v téže oblasti /např. B_3 a B_6 / jsou kladně asociovány, s hodnotou Yuleova $Q=0.1$. Teoretické pravděpodobnosti jednotlivých kombinací hodnot "nula" a "jedna" pro deset veličin jedné oblasti jsou odvozeny z log-lineárního hierarchického modelu

Při generování souborů pro klinické studie náš program generoval náhodně vektory 50 nul a jedniček popsaného teoretického rozdělení tak dlouho, až byl získán požadovaný počet "zdravých" a "nemocných" osob. Tento "náhodný výběr osob" byl jediným stochastickým prvkem celého experimentu.

Studie. Simulovali jsme klinické studie, prosté všech problémů, diskutovaných v úvodu:

- /1/ Veličin, souvisejících se vznikem nemoci je relativně málo a žádnou z nich jsme neopomněli sledovat.
 - /2/ Patogenetický mechanismus je v určitém smyslu jednoduchý: lze ho popsat v několika větách přirozeného jazyka /Srv. kap. 3/.
 - /3/ Patogenetický mechanismus má naprosto deterministický charakter.
 - /4/ Všechny údaje jsou zcela spolehlivé.
 - /5/ Klinický obraz je neměnný v místě a čase.
- V posledních studiích /K2, K3/ byl v řádově rozumném rozsahu odstraněn i poslední diskutovaný problém:
- /6/ Rozsahy souborů přibližně odpovídají řádově největším prakticky realizovatelným studiím.

Přesto se "skutečný" patogenetický mechanismus nepodařilo našim simulovaným pilným, chytrým /a také svárlivým/ výzkumníkům objevit.

Reálný výzkumný pracovník se nikdy nemůže dostat do situace, kdy problémy sub 1/ až sub 6/ neexistují a proto o problémech sub 7/, 8/ ... může nejspíše jen spekulovat. Dr. Ing., dr. Yingová a prof. King však existenci těchto problémů mohou z vlastní zkušenosti potvrdit.

5. DISKUSE A VÁHÁNÍ

- /1/ Je naše studovaná osoba "reálná" nebo "imaginární"? Pravidla, která byla použita k rozhodování zda osoba je nemocná či ne, by mohla být - v zásadě - pravidly některé choroby reálného světa. Taková choroba je možná a v jistém smyslu dokonce "reálná", protože existuje, třebaže "pouze v počítači".
- /2/ Co je jednoduché a co složité. Rozlišujeme tři možné situace:
 - /a/ Lékařský výzkumník má o patogenetickém mechanismu přesnou představu, popsanou v kapitole 3 a žádá statistické ověření své hypotézy. Úloha pro statistika je zde zcela triviální a okamžitě řešitelná.
 - /b/ Výzkumníkovy představy o chorobě vedou k rozumné třídě možných patogenetických mechanismů. Úloha může být s trochou štěstí a větší péčí prakticky řešitelná.
 - /c/ Výzkumník předpokládá existenci patogenetického mechanismu, není však schopen o něm říci něco podstatně konkrétnějšího. To je situace, kterou jsme modelovali: Mechanismus může být relativně jednoduchý a dokonce "blízko" našim představám, nicméně mimo třídu modelů, které jsme schopni předpokládat. Rádi bychom

prorazili bariéru myšlení, nikdo nám ale neporadí, kterým směrem. Toto je velmi těžká "špatně definovaná" úloha /"ill-defined task"/, protože ani analýza dat, ani jiné matematické disciplíny /např. teorie automatů/ nenabízejí algoritmy nebo metody, které by vedly k zaručenému úspěchu.

/3/ Kombinatorický pohled. Z pohledu matematika je základní problém - kromě zmíněné špatně definovatelnosti - v prakticky nekonečné variabilitě formálně možných patogenetických mechanismů. Výzkumníka si můžeme představit jako drobnou postavu, hledající východ z rozlehlého labyrintu. Při pohledu se shora by se jeho cesta mohla jevit jako celkem krátká; reálná cesta, která ho čeká však zřejmě bude velmi dlouhá či prakticky nekonečná. I pokud dojde k svému cíli, vznikne mu problém verifikace, zda skutečně našel "správné" vysvětlení patogenetického mechanismu. Při velkém množství formálně možných patogenetických mechanismů je myslitelné, že více z nich bude v souladu s pozorováními na velkém množství reálných osob, nicméně v souladu náhodně bez exploračního nebo predikčního významu.

/4/ Pohled lékařského výzkumníka. Lékař může v našem experimentu nalézt argumenty i (a) pro klinické stanovisko, že každého konkrétního pacienta s konkrétní chorobou je nutno studovat a pochopit zcela individualizovaně, i (b) pro stanovisko "osvíceného výzkumníka", že při poznávání reality nelze postupovat formálně a mechanicky, ale že je nutno se hojně opírat o fundované úvahy a intuice. /Matematici rádi souhlasí s oběma názory: s prvním jako pacienti, s druhým jako spolupracovníci/.

/5/ Jaké nápady by dostal reálný analytik dat? /Dovětek na poslední chvíli/. psychologického hlediska nebylo možno náš experiment uskutečnit zcela důsledně jen námi: Znajíce skutečný mechanismus nemoci, zda by výzkumník při analýze např. kingenských dat dostal v klíčovém okamžicích nápady, nezbytné k rekonstrukci mechanismu. Proto jsme velmi uvítali pomoc pracovníků Matematického ústavu ČSAV RNDr. Petra Hájka, CSc a RNDr. Kamily Bendové, kteří náš experiment pojali jako výzvu k vyzkoušení heuristické síly metody GUHA. Dr. Bendová, matematicka /obor: matematická logika/ s velkými zkušenostmi v oblasti analýzy reálných dat pomocí procedur metody GUHA, se obeznámila se základními informacemi o experimentu, zhruba v rozsahu 1. a 2. kapitoly tohoto článku. Další informace, zejména definice choroby /kap. 3/ zůstaly před ní velmi pečlivě utajeny. Dr. Bendová věděla pouze to, že jde o arteficiální záležitost a že definice choroby je textově stručná. Dr. Bendové jsme předali dva datové soubory; jeden s 80+80 záznamy a studii Kingen III s 4000+4000 záznamy o "zkoumaných osobách". Je třeba říci, že experiment jsme v žádném případě nekoncepovali jako demonstraci přednosti metody GUHA; naopak, předpokládali jsme, že zrnka pravdy budou dobře ukryta v oceánech zavádějících hypotéz.

Na semináři aplikované matematické logiky /12.2.1986 v Matematickém ústavu ČSAV/ referovala Dr. Bendová své výsledky. Definice, kterou objevila, je zcela ekvivalentní s naší definicí rozdílu mezi zdravými a nemocnými osobami. Definice je stručná, formulovaná mírně odlišným způsobem. Definici Dr. Bendová objevila po velmi důkladné analýze ze souboru Kingen III; analýza menšího souboru sama až k objevu nevedla. Dr. Bendová popsal

chronologický průběh řešení, včetně řady slepých uliček. Rozhodně se nejednálo o mechanickou aplikaci počítačových procedur, rozhodující byly "správné nápady ve správných okamžicích". Diskusi k úspěšnému řešení necháváme zatím otevřenou, experiment Dr. Bendové bude podrobně popsán v samostatné publikaci.

Asi jen další experimenty by ukázaly, nakolik metoda GUHA byla v tomto poznávacím procesu nezastupitelná, zda by úloha byla řešitelná za pomoci standartnějších postupů matematické analýzy a explorační analýzy dat, nebo zda by nebyla řešitelná dokonce strukturálními úvahami nad daty zcela bez použití počítačů.

Jisté je, že mezi naší chorobou - představující obtížný, ale reálně řešitelný problém - a infarktem myokardu nebo jinou reálně existující multifaktoriálně podmíněnou chorobou je velké rozpětí. Bylo by zajímavé modifikovat náš experiment tak, že bychom postupně různým způsobem komplikovali jednotlivé podmínky (1) a (5), resp. (6): např. bychom definovali chorobu kombinatoricky jen o trochu složitěji /podmínka (2)/. Pomocí těchto experimentů bychom zkoušeli hledat, kudy prochází hranice reálně řešitelnosti. Pro tyto experimenty nemáme dostatek pokusných osob.